////// 高级生信系列课程 //////

利用Python进行生信分析

原理 · 算法 · 实战

编程课

高级生信系列课程

项目二
BED文件的操作与处理

孟浩巍
menghaowei1992@126.com
2022年9月25日

生物信息学日常学习的痛点

- 常见的几种路径
  - 基于ctrl + c / ctrl + v 的学习
  - “照猫画虎”式学习
  - “百家饭”式学习
  - “继承遗产”式学习

- 痛点：
  - 遇到问题不知道用什么；
  - 遇到问题不知道怎么用；
  - 遇到bug不知道怎么解决bug；

课程资料下载

- 由于腾讯课堂改版，所有课程讲义、测试数据、运行代码均**需要加入课程讨论群**进行下载！

- 课程讨论QQ群号： 828030346

本节课程授课顺序

1. 讲解相关背景；
2. 讨论程序编写的逻辑；
3. 使用jupyter编写程序，并调试程序；
4. 将程序编写为命令行形式；

Part I BED文件与bedtools

什么是BED格式?

1. 文本文件;
2. 表征基因组的一段区域;
3. 标准的BED文件最少3列，最多12列;

BED格式每列信息详解

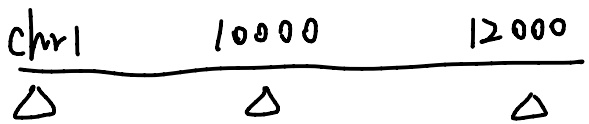
1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature, however, the number in **position format** will be represented. For example, the first 100 bases of chromosome 1 are defined as *chrom=1, chromStart=0, chromEnd=100*, and span the bases numbered 0-99 in our software (not 0-100), but will represent the position notation chr1:1-100. Read more **here**.
   chromStart and *chromEnd* can be identical, creating a feature of length 0, commonly used for insertions. For example, use *chromStart=0, chromEnd=0* to represent an insertion before the first nucleotide of a chromosome.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade									
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

6. **strand** - Defines the strand. Either "." (=no strand) or "+" or "-".
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.



使用IGV检查不同详细程度BED格式的展示

1. 包含前3列信息的行；（chrom, start, end）
2. 包含前4列信息的行；（chrom, start, end, name）
3. 包含前5列信息的行；（chrom, start, end, name, score）
4. 包含前6列信息的行；（chrom, start, end, name, score, strand）

BED格式的常用举例

1. 储存基因区；
2. 储存基因组的某些位点信息，如TSS等；
3. 储存ChIP-seq，ATAC-seq，CUT&Tag等的富集peak信息；

…… ……

BED文件的操作与bedtools

- bedtools是一种常用的bed操作工具，可以实现非常多的常用功能
  - 2个bed文件的交集（bedtools intersect）；
  - bed文件按照基因组坐标排序（bedtools sort）；
  - 对bed文件进行扩大，平移（bedtools shift）；
  - 对bed文件进行随机抽取（bedtools random）；
  - 根据提供的bed文件在基因组进行随机抽样（bedtools shuffle）
  - 计算bed文件区域内比对到的reads数（bedtools coverage）
  - ……

测试数据的说明

- ENCODE数据集
- https://www.encodeproject.org/experiments/ENCSR617IFZ/

- CTCF rep1 BED
- https://www.encodeproject.org/files/ENCFF684NDR/@@download/ENCFF684NDR.bed.gz

- CTCF rep2 BED
- https://www.encodeproject.org/files/ENCFF151CRB/@@download/ENCFF151CRB.bed.gz

- 参考基因组下载
- https://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/

操作的用途举例

- 2个bed文件的交集（bedtools intersect）；
  - 输入文件 BED a， BED b

https://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html

操作的用途举例

- bed文件按照基因组坐标排序（bedtools sort）；
  - 输入文件 BED，基因组长度文件 .fai

- 对bed文件进行扩大，平移 （bedtools shift）；
  - 输入文件 BED

https://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html

操作的用途举例

- 对bed文件进行随机抽取 （bedtools random）；
  - 输入文件 BED

- 根据提供的bed文件在基因组进行随机抽样 （bedtools shuffle）
  - 输入文件 BED，基因组长度文件 .fai

https://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html

操作的用途举例

- 计算bed文件区域内比对到的reads数 （bedtools coverage）
  - 输入文件：BED，BAM

https://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html

为什么已经有了工具我们还要自己写程序?

1. 学习编程技巧;
2. 深刻理解工具的参数、用法、设计思路;
3. 触类旁通，举一反三

Part II
获取BED文件的交集与并集

程序分析（2个BED文件求交集）

- 程序的输入

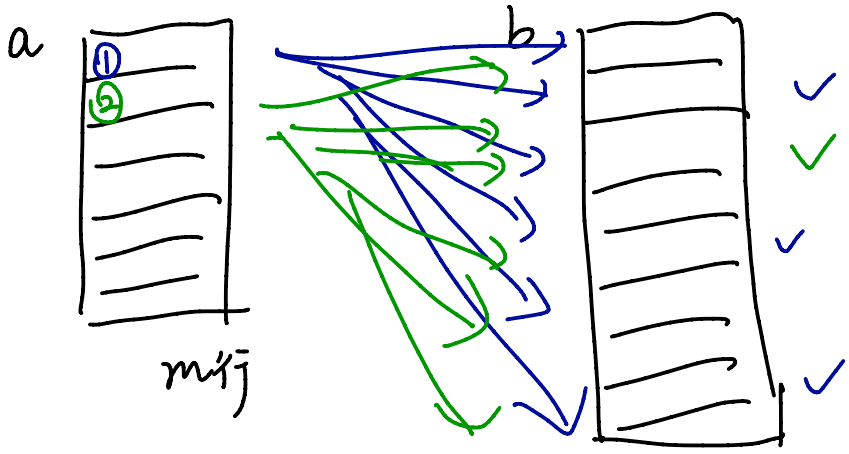
  BED a , BED b

- 程序的输出

  输出a, count, region, sum(region.length) / total length .

- 程序的用途
  - 求解交集在a文件中的情况；
  - 求解交集的部分；
  - 求解交集部分的占比；
  - 其它个性化需求……

思路1：直接进行全部循环



m+n次

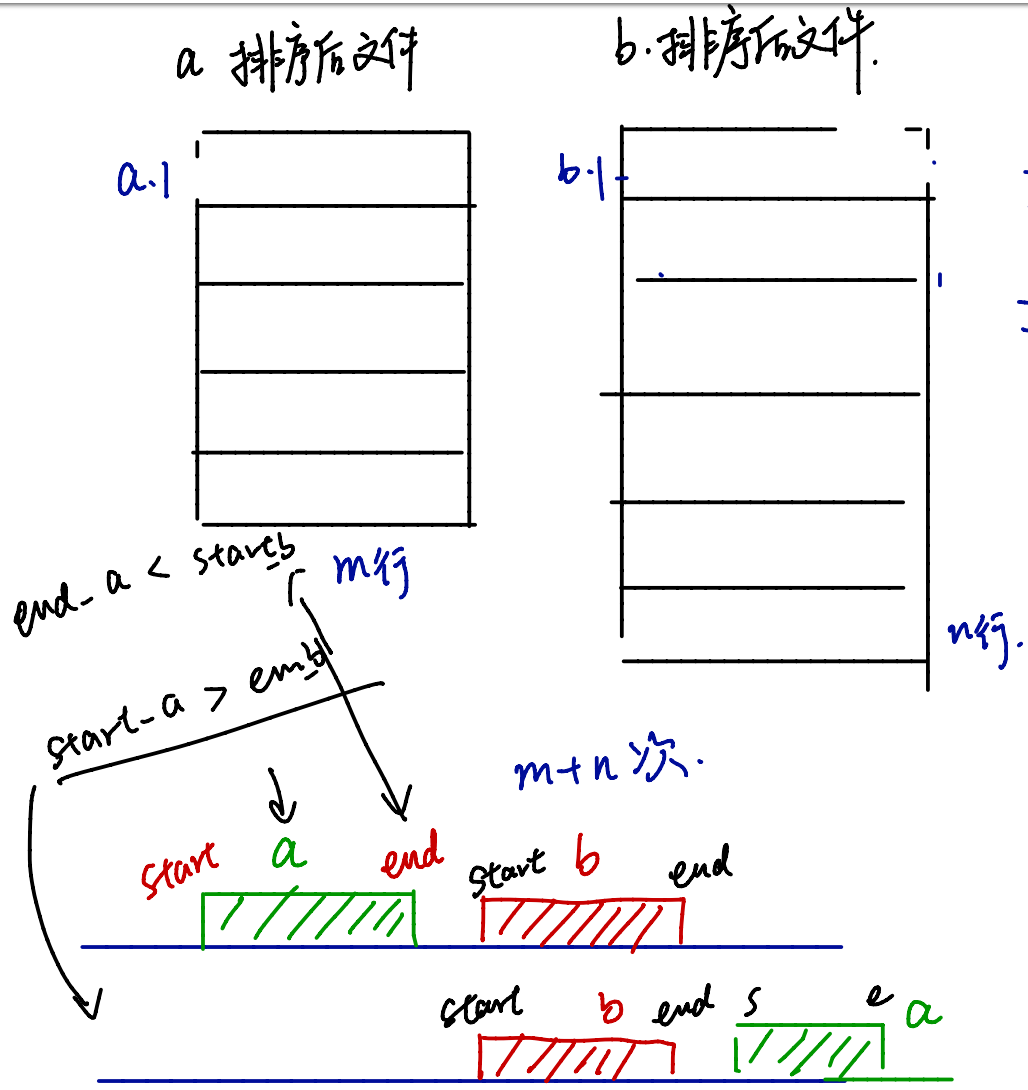
m*n次.

300000 * 250000

= 750 10^8

= 7.5×10^{10}

思路2：根据排序后的BED文件进行操作



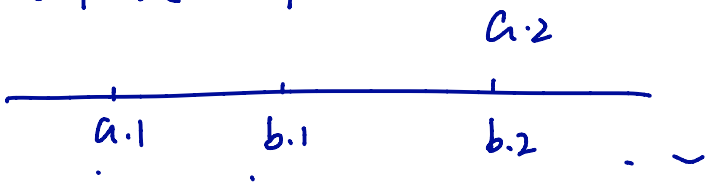
a.1 vs b.1

I. 有交叉，有交集 → 输出.

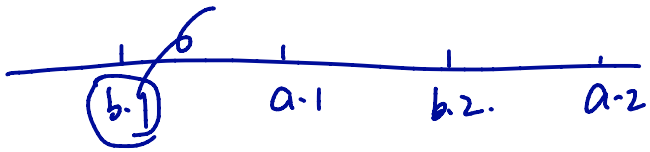
II. 无交集

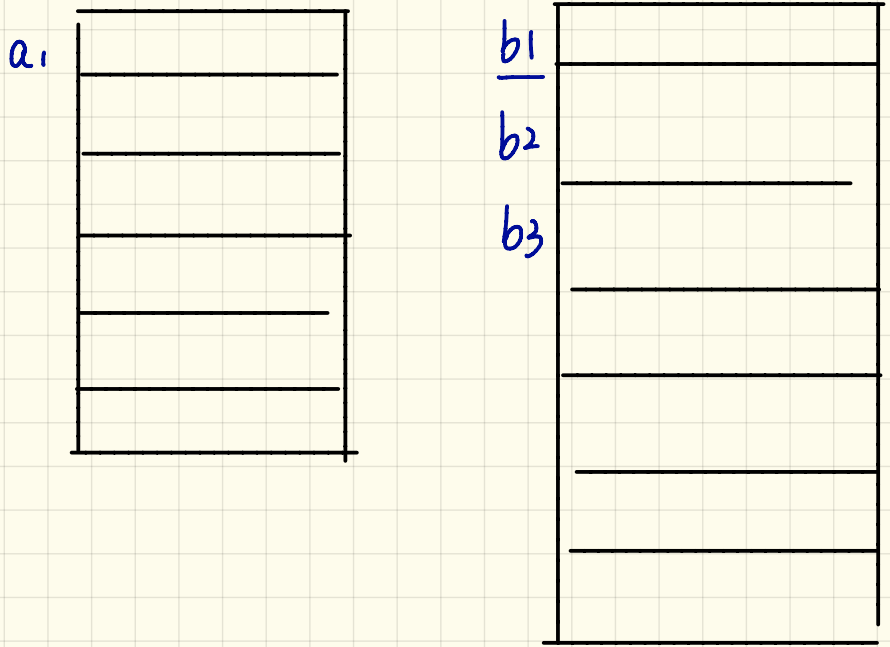
① a.1 比 b.1 靠上游.

a.1 一定无b中交集. a.1 扔掉.



② a.1 比 b.1 靠下游.




a1
b1
b2
b3

程序分析（2个BED文件求并集）

- 程序的输入
- 程序的输出
- 程序的分析
  - 理论上，交集能搞定，并集就能搞定

程序思路：求两个BED的并集

////// 高级生信系列课程 //////

1kbp bin,
2kbp. bin
1Mb bin

分bin.

Part III
将基因组分成等长区间并计算CG含量

程序分析（2个BED文件求交集）

- 程序的输入

基因组长度. 基因组序列

- 程序的输出

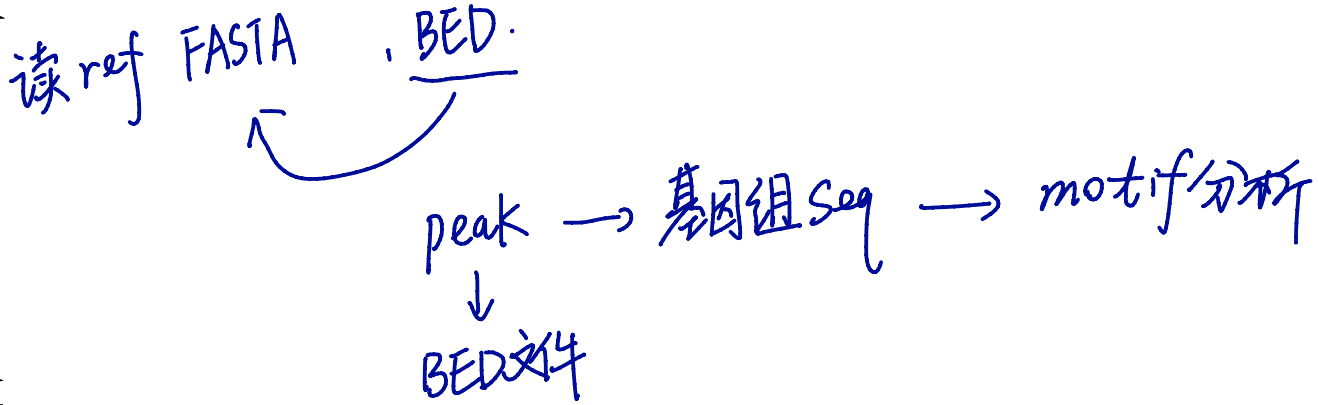
- 程序的用途

程序设计思路

Part IV
根据给定的BED文件提取基因组序列

程序分析（2个BED文件求交集）

- 程序的输入
- 程序的输出
- 程序的用途



程序设计思路

Thanks for your attention!
And see you next time!

Haowei MENG
menghaowei@pku.edu.cn
2022-09-24